DOCUMENT RESUME

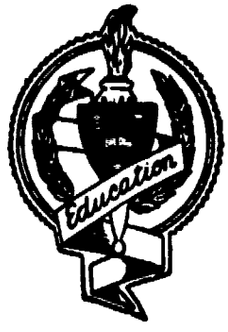
ED 080 312 SE 016 110

AUTHOR Aikenhead, Glen S.
TITLE A New Methodology for Test Construction in Course Evaluation.
INSTITUTION Saskatchewan Univ., Saskatoon. Dept. of Curriculum Studies.
PUB DATE Mar 73
NOTE 14p.; Paper presented at the annual meeting of the National Association for Research in Science Teaching (46th, Detroit, Michigan, March 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Educational Research; *Evaluation; *Evaluation Techniques; *Physics; Science Education; Secondary School Science; *Test Construction; *Testing
IDENTIFIERS Research Reports

ABSTRACT

A new method of constructing tests is presented in this article for the purpose of developing a test from student perception of the course. The Test on Understanding Science (TOUS) and the Science Process Inventory (SPI) were used as sources of items. A random sub-sample of 921 students, taking both the pretest and posttest of TOUS and SPI during the 1967-68 Project Physics (PP) experimental period, served as sources of empirical data. The McNemar chi square analysis was used to select test items empirically. Every item was analyzed with respect to the changes in student responses between the pretest and posttest. The items showing a statistically significant change in response were combined into a single instrument called "A Measurement of Knowledge About Science and Scientists (Project Physics: Form 1)" (KASSPP1). Another independent random sub-sample of 64 students was tested to describe the statistical attributes of KASSPP1. Findings showed that KASSPP1 had a greater predictive validity for PP than either TOUS or SPI. Application of the present method to formative evaluation was recommended. (CC)



Information and Research Report

Department of Curriculum Studies

College of Education

University of Saskatchewan

Saskatoon, Canada

A NEW METHODOLOGY FOR TEST CONSTRUCTION IN COURSE EVALUATION

Dr. Glen S. Aikenhead
Assistant Professor
Science Education

Paper presented at the 46th annual meeting of the
National Association for Research in Science Teaching
DETROIT, MICHIGAN
March 26 - 29, 1973

INTRODUCTION

Test validity has always been a prime concern to an evaluator. Test developers carefully construct or select test items that yield greatest validity because this quality in an instrument lends credence to the results that emerge from its use. This report presents a new method in test construction which augments an instrument's validity beyond that possible by standard procedures.

VALIDITY

A short delineation of validity serves to clarify the issues discussed in this report. Validity, "the degree to which a test is capable of achieving certain aims," may be characterized by the following four categories: content, predictive construct, and concurrent validity.[1]

Content validity simply refers to the substance or content of a test's being representative of the content of the property being measured. Often, content validity relies on the consistency between test items and the writings of respected authors. Models and outlines of a subject field and "expert" judgements are usually utilized for this purpose.

Predictive validity concerns an instrument's ability to predict certain observable performances. That is, from a statement of what a test claims to measure and from a person's score, one can infer or predict certain behavior. For instance, the College Board Entrance Examinations predict to some degree a student's success in his first year of university. In addition, an instrument would be thought to have high predictive validity if it would distinguish between students who had undergone a certain treatment and students who had not; for example, a test purported to measure a student's appreciation of poetry would have high predictive validity if it could identify those students who had taken a course in poetry appreciation. A mathematics test would have no predictive validity if it could not discriminate between persons who have studied the subject extensively and those who have not.

A test's construct validity is determined by knowing what factors (psychological properties) "explain" the variance in the scores of that test. The constructs in construct validity refer to the attributes that are supposedly reflected in the test performance. Factor analysis and canonical correlation are tools in determining construct validity.[2] For example, a measure of IQ is a popular psychological property and researchers are often interested in knowing to what extent IQ scores "explain" the variance in the scores of a certain instrument.

Concurrent validity refers to the degree of consistency among scores of similar tests. That is, a new instrument may be thought to be valid if it is shown to yield similar results as an established instrument. A factor analysis with such a reference test is ideal.

TEST CONSTRUCTION METHODOLOGY

Standard procedures in test construction include formulating a large set of items which appear to test a student's grasp of an idea, concept or relationship in a given subject area (the learning one hopes will occur). This is believed to establish content validity. After a trial run, items are eliminated for various reasons:

(a) poor or ambiguous wording leading to low reliability,

(b) low "point biserial correlation" (a measure of an individual item's ability to differentiate between a student scoring high on the total test and a student scoring low on the total test). It is thought that items that differentiate to a high degree are useful in augmenting the range of total test scores; that is, increase the ease for ranking students. For instance, if there are items for which "poor" students invariably do well whereas "better" students do not, then these items would be deleted from the test. Also, items that are too difficult and too easy would have low point biserial correlations.

(c) closely associated with (b), inability to discriminate between "experts" and "non-experts" in the field being tested.

The "good" items are combined into a test and subjected to further trials. The elimination process increases the predictive validity of the instrument. Further analysis would be necessary to establish construct and concurrent validity if this were considered desirable.

The content of each item of the test corresponds to what the test writer believes to be the course objectives. That is, each item reflects the course content as seen through the eyes of the test constructor. Gains in student achievement on these instruments

only coincide with what the teacher or curriculum developer HOPES students learn. If students do not gain significantly, one assumes they have not learned sufficiently. But what is the course content as seen through the eyes of the student? If students do not gain significantly, then the hoped-for learning did not take place because it was not adequately provided.

In evaluating a course, one should ask: What do students generally learn? Only after further analysis would course objectives be considered: Which objectives were accomplished and which were not? What learning - positive and negative - took place that was not encompassed by the course objectives? (A question all too few evaluators consider).

The standard procedure for test construction narrowly deals with hoped-for learning based on a fragile correspondence among course objectives, course content, student experience in that course, and the test items' content. An alternative procedure has been developed that tends to overcome these limitations in the standard method of test construction.

AN ALTERNATIVE PARADIGM

I propose a new procedure for test development. This alternative paradigm is quite simple: a test is constructed from the students' perception of the course, instead of the teacher's or curriculum developer's vision of the course. This new methodology gives high predictive validity to the evaluative instrument and, as demonstrated shortly, yields greater feedback to the teacher,

student, and curriculum developer.

To develop a specific test for a given course, one empirically selects test items from any number of sources, including existing validated instruments. (These items are broader in scope than the individual course objectives, assuring high content validity). The empirical basis of this selection rests upon the demonstrated achievement (not just the hoped-for achievement) of a large number of students studying the particular course. The items chosen are those which show a statistically significant change in student response between a pretest and posttest administration of the original instruments. These are items on which students make a significant improvement, or on which students make a significant decrease, in their number of correct responses. Davis[3] contends that "the most important objective of evaluation in education is to estimate changes in individual learners and groups of learners."

The proposed paradigm in test construction yields a new type of test which is comprised of items chosen for their ability to indicate changes in learners who have studied a certain course. The derived instruments would have greater predictive validity for that particular course than any of the original instruments. This increased predictive validity would tend to yield a greater amount of feedback to both teacher and student because the derived test would tend to show more change in student knowledge than the parent instruments.

The empirical nature of the item selection emerges from a statistical item analysis. Every item is analyzed with respect to the changes in student responses between the pretest and posttest. There are only two changes possible: (a) from an incorrect response to correct response, and (b) from a correct response to an incorrect response. The probability is .50 that students who change their response move to a correct answer. McNemar[4] has derived a chi square test specifically for this situation:

$$\chi^2 = \frac{(A - D)^2}{(A + D)}$$

with 1 degree of freedom*, where A and D are cell frequencies in the following contingency table:

Item X

Pretest \ Posttest	1	0
0	A	B
1	C	D

1 signifies a correct response

0 signifies an incorrect response

The chi square analysis determines which items experience a statistically significant change in student response between the pretest and posttest. These items may be combined into a single instrument, a test specifically applicable to a given course. In addition, one may infer what knowledge students appeared to gain or lose during the interim.

---

*When A+D < 20, a Yate's correction for small frequencies must be introduced. The equation becomes: $\chi^2 = (|A-D| - 1)^2 / (A+D)$.

FIELD TRIAL

Description

In an experimental study, this new methodology for constructing tests successfully led to a unique evaluation instrument for the new physics course by Holton, Rutherford and Watson, Harvard Project Physics (HPP).* The investigation used two validated instruments, the Test on Understanding Science (TOUS)[5] and the Science Process Inventory (SPI)[6], as the original source of items. Many of the HPP's major objectives appear to be encompassed by the content of the TOUS and SPI. The HPP student responses to the TOUS and SPI items (pretest and posttest) yielded the empirical data for this study.

The students were part of a nation-wide evaluation of HPP. Fifty-five teachers were selected at random from the population of American and Canadian physics teachers.[7] These randomly selected teachers were again randomly split into two groups: thirty-five taught HPP while twenty served as a control group teaching their usual physics courses (non-HPP). The HPP group attended a summer institute to prepare them to teach the new course. In addition, there was another group of teachers experienced in teaching HPP. These nineteen had volunteered to participate in the evaluation project. They taught in various regions of the United States. The number of students studying HPP in the evaluation project totalled 2,950. From this group, 921 students were randomly chosen to write both the pretest and posttest TOUS or SPI.[8]

---

*The first commercial edition was published as Project Physics, September, 1970, Holt, Rinehart & Winston.

Similarly, sixty-four students were randomly selected to write both tests both times. (The HPP evaluation project used many other instruments. This randomization reduced the total time taken by testing.[9])

Results

The TOUS and SPI supplied 195 items of which 101 items were selected by the observed significant changes in student knowledge over a year of studying HPP. This derived HPP test was called "A Measurement of Knowledge About Science and Scientists (Project Physics: Form 1)", abbreviated KASSPP1. The test includes ninety-five items which showed a significant positive gain and six items which experienced a significant negative gain in student response. The rationale for including these six items is presented below.

Some quantitative attributes of the KASSPP1 may be found in Table I.

TABLE I

QUANTITATIVE DATA FOR THE KASSPP1
BASED ON THE RANDOM SUBSAMPLE OF HPP STUDENTS

	Mean Score	SD[a]	Range	Pre-post Correlation	Reliability[b] estimate	N
Pretest	70.70	7.403	52 - 82	.76	.79	64
Posttest	78.16	8.663	53 - 95			

[a] S D means standard deviation.

[b] K u d e r Richardson formula-20 was used.

These data were obtained from the random subsample of sixty-four HPP students. The mean scores lay half way between a score obtainable by pure chance (47 points) and a perfect score (101 points). The standard deviation (7.403 on the pretest) was similar to the TOUS (7.13)[10] but was not as large as the standard deviation of the SPI (13.1)[11]. The range of scores showed a spread of 30 to 40 points. Two different estimates of the test's reliability compare favorably with the standards established by Davis[12] for measuring group and individual characteristics. The relationship between KASSPP1 scores and measures of reading and "mental" ability is assumed to fall within the range set by the TOUS and SPI (correlation coefficients of .47 to .66).[13]

In constructing the KASSPP1, primary consideration was given to its ability to reflect changes in student knowledge, and not to maximizing its quantitative attributes. Thus, items were included that showed a negative change between the pretest and posttest. If the number of items experiencing negative gains was relatively small, one would expect the KASSPP1 to yield larger gain scores than either the TOUS or SPI. This expectation was fulfilled by the results of the independent analysis of a random subsample of HPP students. This subsample's KASSPP1 gain score (7.46 points) exceeded its TOUS gain score (2.76 points) and its SPI gain (5.80 points). These results documented the increased predictive validity of the KASSPP1. With regard to the number of items that

experienced a significant change between the pretest and posttest, the KASSPP1 also yielded more information than either the TOUS or SPI. The random subsample of HPP students demonstrated a very large improvement for 50 TOUS and SPI items, 26% of all 195 items. Equally large gains were accomplished for 38 KASSPP1 items, 38% of the total 101 items. Thus, the one test (KASSPP1) was able to yield a somewhat higher proportion (38% compared wi[illegible] items on which students dramatically moved toward the correct response. This result suggests that the KASSPP1 is more efficient than the TOUS and SPI in yielding feedback for HPP teachers and students.

IMPLICATIONS FOR FURTHER RESEARCH

The development of the KASSPP1 illustrates a novel method of test construction: general and valid instruments are utilized by empirically selecting items which prove to be most appropriate to a particular course. Many studies may be conducted by employing this new paradigm. For any curriculum in its early stages of development the following procedures are suggested.

(a) General tests, such as the TOUS and SPI, could be used to observe what knowledge students tended to learn when studying a new course. The information may lead to alterations or shifts in emphasis in the curriculum materials.

(b) The testing would then be repeated for the last revision of the curriculum. Tests applicable to the new course could then be derived.

(c) The curriculum project's package of evaluation instruments would include these derived tests. This is especially useful in the case where the derived instruments concern knowledge traditionally thought to lie beyond the realm of standard subject matter; for instance, knowledge closely related to students' impressions and attitudes.

Not only did the new method of item selection yield an objective test about science and scientists, it also supplied sufficient data for partially evaluating the HPP course.[14] By examining the items that experienced significant gains or losses, one can recognize differences between HPP and other physics courses. These differences were defined in terms of the knowledge students tended to acquire rather than in terms of differences in student mean scores. Student achievement was also compared with the objectives of HPP.[8] Such analyses and comparisons correspond to major components of formative evaluation.

REFERENCES

1. Technical Recommendations for Psychological Tests and Diagnostic Techniques. Psychological Bulletin, 51 (Supplement, 1954), 13.

2. Mayo, Samuel T. "The Methodology and Technology of Educational and Psychological Testing." Review of Educational Research, 38 (February, 1968), 92-101.

3. Davis, Frederick B. "Testing and the Use of Test Results." Review of Educational Research, 32 (February, 1962), 5-14.

4. McNemar, Quinn. Psychological Statistics. 4th ed. New York: John Wiley and Sons, 1969.

5. Cooley, William W., and Klopfer, Leo. Test on Understanding Science. Princeton, New Jersey: Educational Testing Service, 1961.

6. Welch, Wayne W. Welch Science Process Inventory, form D. Cambridge, Mass.: Harvard Project Physics, 1966.

7. Welch, Wayne W; Walbert, Herbert J.; and Ahlgren, Andrew. "The Selection of a National Random Sample of Teachers for Experimental Curriculum Evlauation." School Science and Mathematics, 49 (March, 1969), 210-216.

8. Aikenhead, Glen S. "The Measurement of Knowledge About Science and Scientists: An Investigation into the Development of Instruments for Formative Evaluation." Unpublished Doctoral Thesis, Harvard University, 1972. p. 91 & pp. 181-188.

9. Walberg, Herbert J. and Welch, Wayne W. "A New Use of Randomization in Experimental Curriculum." School Review (Winter, 1967).

10. Cooley, William W., and Klopfer, Leo. Manual: Test on Understanding Science. Princeton, New Jersey: Educational Testing Service, 1961.

11. Welch, Wayne W. "Welch Science Process Inventory, Form D: Summary of Information." 104 Burton Hall, University of Minnesota, Minneapolis, Minn.: Dr. Wayne W. Welch. (Mimeographed, no date listed.)

12. Davis, Frederick B. Educational Measurements and their Interpretations. Belmont, California: Wadsworth, 1964.

13. Aikenhead, Glen S. "The Measurement of High School Students' Knowledge About Science and Scientists." Unpublished Qualifying Paper, Harvard University, 1970.

14. Aikenhead, Glen S. "The Interpretation of Student Performance on Evaluative Tests." Paper presented at the 46th annual meeting of the National Association for Research in Science Teaching, Detroit, March 26-29, 1973.